043290.P8795 *Patent*

UNITED STATES PATENT APPLICATION

for

SYSTEM AND METHOD FOR FAULT TOLERANT STREAM SPLITTING

INVENTOR

Reed Sloss

Prepared by:

BLAKELY, SOKOLOFF, TAYLOR & ZAFMAN, LLP
12400 WILSHIRE BOULEVARD
SEVENTH FLOOR
LOS ANGELES, CALIFORNIA 90025
(408) 720-8598

Attorney's Docket No. 042390.P8795

"Express Mail" mailing label number EL034159856US

Date of Deposit June 30, 2000

I hereby certify that this paper or fee is being deposited with the United States Postal Service "Express Mail Post Office to Addressee" service under 37 C.F.R. 1.10 on the date indicated above and is addressed to the Commissioner of Patents and Trademarks, Washington, D.C. 20231.

Cindy Murphy
(typed or printed name of person mailing paper or fee)

Cindy Murphy
(Signature of person mailing paper or fee)

SYSTEM AND METHOD FOR FAULT TOLERANT STREAM SPLITTING

BACKGROUND OF THE INVENTION

Field of the Invention

This invention relates generally to the field of network services. More particularly, the invention relates to an improved architecture for providing fault tolerant data communication.

Description of the Related Art

As is known in the art, streaming is a mechanism for playing back audio and/or video content over a network in real-time, typically used in situations where network bandwidth is limited. The basic streaming concept is that the destination (e.g., a client) begins to play back the underlying streaming file from a buffer before the entire file has been received from its source.

A traditional network streaming system is illustrated in **Figure 1**. As shown, one or more clients 150, 160, configured with streaming application software such as RealPlayer ® from RealNetworks® or Windows Media® Player from Microsoft® Corporation, communicate with one or more streaming servers 110, 111, . . . N, over a network 100 (e.g., the Internet). The group of streaming servers 110, 111, . . . N, are located together at a point of presence ("POP") site. Each of the streaming servers 110, 111, . . . N, may store a copy of the same streaming data or, alternatively, may store different streaming data, depending on the configuration at the POP site 130.

sub A3 In operation, when a client 150 requests a particular streaming file from a server at the POP site 120, the request is received by a load balancer module 120, which routes the request to an appropriate streaming server 111. Which server is "appropriate" may depend on where the requested file is stored, the load on each server 110, 111, . . . N, and/or the type of streaming file requested by the client (e.g., Windows Media format or RealPlayer format). Once the file has been identified by the load balancer 120 on an appropriate server – server 111 in the illustrated example – it is streamed to the requesting client 150 (represented by stream 140) through the network 100.

BRIEF DESCRIPTION OF THE DRAWINGS

A better understanding of the present invention can be obtained from the following detailed description in conjunction with the following drawings, in which:

FIG. 1 illustrates a prior art system and method for streaming content over a network.

FIG. 2 illustrates an exemplary network architecture including elements of the invention.

FIG. 3 illustrates an exemplary computer architecture including elements of the invention.

FIG. 4 illustrates various streaming sources supported by one embodiment of the invention.

FIG. 5 illustrates stream splitting implemented in on embodiment of the invention.

FIG. 6 illustrates a system for implementing a backup root splitter according to one embodiment of the invention.

FIG. 7 illustrates a method for implementing a backup root splitter according to one embodiment of the invention.

DETAILED DESCRIPTION

In the following description, for the purposes of explanation, numerous specific details are set forth in order to provide a thorough understanding of the present invention. It will be apparent, however, to one skilled in the art that the present invention may be practiced without some of these specific details. In other instances, well-known structures and devices are shown in block diagram form to avoid obscuring the underlying principles of the invention.

AN EXEMPLARY NETWORK ARCHITECTURE

Elements of the present invention may be included within a multi-tiered networking architecture 200 such as that illustrated in **Figure 2**, which includes one or more data centers 220-222, a plurality of "intermediate" Point of Presence ("POP") nodes 230-234 (also referred to herein as "Private Network Access Points," or "P-NAPs"), and a plurality of "edge" POP nodes 240-245 (also referred to herein as "Internet Service Provider Co-Location" sites or "ISP Co-Lo" sites).

According to the embodiment depicted in **Figure 2**, each of the data centers 220-222, intermediate POPs 230-234 and/or edge POPs 240-245 are comprised of groups of network servers on which various types of network content may be stored and transmitted to end users 250, including, for example, Web pages, network news data, e-mail data, File Transfer Protocol ("FTP") files, and live & on-demand multimedia streaming files. It should be noted, however,

that the underlying principles of the invention may be practiced using a variety of different types of network content.

Sub A4 The servers located at the data centers 220-222 and POPs 230-234; 240-245 may communicate with one another and with end users 150 using a variety of communication channels, including, for example, Digital Signal ("DS") channels (e.g., DS-3/T-3, DS-1/T1), Synchronous Optical Network ("SONET") channels (e.g., OC-3/STS-3), Integrated Services Digital Network ("ISDN") channels, Digital Subscriber Line ("DSL") channels, cable modem channels and a variety of wireless communication channels including satellite broadcast and cellular.

In addition, various networking protocols may be used to implement aspects of the system including, for example, the Asynchronous Transfer Mode ("ATM"), Ethernet, and Token Ring (at the data-link level); as well as Transmission Control Protocol/Internet Protocol ("TCP/IP"), Internetwork Packet Exchange ("IPX"), AppleTalk and DECnet (at the network/transport level). It should be noted, however, that the principles of the invention are not limited to any particular communication channel or protocol.

In one embodiment, a database for storing information relating to distributed network content is maintained on servers at the data centers 220-222 (and possibly also at the POP nodes 230-234; 240-245). This information may include the different POP sites which currently contain a copy of network

content. The database in one embodiment is a distributed database (i.e., spread across multiple servers) and may run an instance of a Relational Database Management System (RDBMS), such as Microsoft™ SQL-Server, Oracle™ or the like.

AN EXEMPLARY COMPUTER ARCHITECTURE

Having briefly described an exemplary network architecture which employs various elements of the present invention, a computer system 300 representing exemplary clients and servers for implementing elements of the present invention will now be described with reference to **Figure 3**.

One embodiment of computer system 300 comprises a system bus 320 for communicating information, and a processor 310 coupled to bus 320 for processing information. The computer system 300 further comprises a random access memory (RAM) or other dynamic storage device 325 (referred to herein as "main memory"), coupled to bus 320 for storing information and instructions to be executed by processor 310. Main memory 325 also may be used for storing temporary variables or other intermediate information during execution of instructions by processor 310. Computer system 300 also may include a read only memory ("ROM") and/or other static storage device 326 coupled to bus 320 for storing static information and instructions used by processor 310.

A data storage device 327 such as a magnetic disk or optical disc and its corresponding drive may also be coupled to computer system 300 for storing information and instructions. The computer system 300 can also be coupled to a second I/O bus 350 via an I/O interface 330. A plurality of I/O devices may be coupled to I/O bus 350, including a display device 343, and/or an input device (e.g., an alphanumeric input device 342 and/or a cursor control device 341).

The communication device 340 is used for accessing other computers (servers or clients) via a network 100. The communication device 340 may comprise a modem, a network interface card, or other well known interface device, such as those used for coupling to Ethernet, token ring, or other types of computer networks.

EMBODIMENTS OF THE INVENTION

Stream Splitting

Embodiments of a system configured to stream live and on-demand audio/video content will now be described with respect to **Figures 4** and **5**. As shown in **Figure 4**, one embodiment receives and processes incoming audio/video content from a variety of sources including, but not limited to, live or recorded signals 401 broadcast over satellite links 410; live signals 402 provided via video conferencing systems 411; and/or live or recorded signals 403 transmitted over dedicated Internet Protocol ("IP") links 412. It should be noted, however, that a various other network protocols (i.e., other than IP) may

be employed while still complying with the underlying principles of the invention. In one embodiment, each of the modules illustrated in **Figure 4** reside at a data center 220.

System acquisition and management modules ("SAMs") 420 open and close communication sessions between the various sources 401-403 as required. For example, when a content provider wants to establish a new live streaming session, the SAM 420 will open a new connection to handle the incoming audio/video data (e.g., after determining that the content provider has the right to establish the connection).

The SAM module 420 will handle incoming signals differently based on whether the signals have already been encoded (e.g., by the content providers) and/or based on whether the signals are comprised of "live" or "on demand" content. For example, if a signal has not already been encoded by a content provider (e.g., the signal may be received at the data center 220 in an analog format or in a non-streaming digital format), the SAM module 420 will direct the signal to one or more streaming encoder modules 1430, which will encode the stream in a specified digital streaming format (e.g., Windows Media® format, Real G2™ format, . . . etc).

If the incoming signal is live, the streaming encoders 430 transmit the encoded signal directly to one or more streaming origin servers 510 (which

distribute the signal to various POP nodes as described below) and/or to one or more content storage devices 431 at the data center 220. If, however, the incoming signal is an on-demand signal (i.e., to be stored and viewed by clients at any time), then the streaming encoders 430 transmit the encoded signal directly to the content storage devices 431. Similarly, if the incoming signal is already encoded in a particular streaming format, it may be transmitted directly to the content storage devices 431.

As new audio/video streaming content is added to the content storage devices 431, the SAM module 420 causes a storage database 430 to be updated accordingly (e.g., via a content delivery subsystem). The storage database 430 in one embodiment is a distributed database which tracks all network content as it distributed and stored at various POP sites throughout the network.

As illustrated in **Figure 5**, the encoded signal is transmitted from the streaming origin servers 510 to streaming splitters 520-522, 530-532 located at a various I-POP nodes 230-232 and E-POP nodes 240-242. Employing streaming splitters as illustrated conserves a substantial amount of network bandwidth. For example, in the illustrated embodiment each streaming splitter 520-522, 530-532 receives only a single stream of live audio/video content from an upstream server, which it then divides into several independent streams. This configuration is particularly useful for streaming configurations which do not support multicasting.

Moreover, employing streaming splitters within a multi-tiered hierarchy, as illustrated in **Figure 5**, reduces bandwidth at each level in the hierarchy. For example, a single stream from a live streaming event may be transmitted from a streaming origin server 510 to an I-POP streaming splitter 521. The streaming splitter 521 may then transmit a single stream to each of the E-POP streaming splitters 530-532, which may then transmit the live event to a plurality of end users 540-548. Accordingly, the network path between the data center 220 and the I-POP 231 is loaded with only a single stream and each of the three network paths between the I-POP 231 and the E-POPs 240-242 are loaded with only a single stream. The incoming streams are then split at each of the E-POPs 240-242 to provide the live event to a plurality of end users 540-548.

Fault Tolerant Stream Splitting

As illustrated in **Figure 6**, one embodiment of the invention includes one or more root splitters 630 configured at POP sites 620 to receive a stream from an origin server and distribute the stream to a plurality of leaf splitters 631-635. Each of the leaf splitters 631-635 serve a plurality of end users 650 by further splitting each stream received from the root splitter 630 into another plurality of end-user streams.

It can be seen from **Figure 6** that, for a particular live or scheduled streaming event, the streaming encoder 530, the origin servers 510 and the root splitter 630 all represent single points of failure. In one embodiment, potential

failures at the data center 200 components (i.e., the origin server 510 and streaming encoder 530) are handled through allocation of redundant encoder/origin server pair combinations for each live/scheduled event. However, in some circumstances, this level of redundancy may be impractical at various POP sites 620 (e.g., due to limited media server resources at the site, limited and costly rack space, . . . etc). As such, in one embodiment, a more efficient mechanism may be implemented to provide fault tolerance at these POP sites 620.

As illustrated in **Figure 6**, in one embodiment, one or more of the leaf splitters (e.g., leaf splitter 631) are configured as backups to the primary root splitter 630. In this embodiment, the health of the root splitter is continually monitored by a monitoring subsystem which may reside on the load balancer module 625, the redirection subsystem 625, or as a separate monitoring module and the data center and/or the POP site 620.

In one embodiment, the root splitter 630 is configured to provide an update to the monitoring subsystem at predetermined intervals. This may be accomplished by an agent 640 continually running on the root splitter 630and configured to communicate with the monitoring subsystem. The periodic update in this embodiment acts as a "heartbeat" which indicates to the monitoring subsystem that the root splitter is operating within normal parameters. If the monitoring subsystem does not receive an update for one or more periods, it

may determine that the root splitter has become inoperative and assign the backup root splitter 631 as the new primary root splitter. In one embodiment, the agent 641 running on the backup root splitter 631 performs the reconfiguration process. Alternatively, or in addition, the monitoring subsystem may actively poll the agent 640 running on the root splitter 630 to verify that the root splitter 630 is operating reliably.

The operation of one embodiment of the system illustrated in **Figure 6** will now be described with respect to the flowchart in **Figure 7**. At 710, a user attempts to view a particular live or scheduled streaming event (e.g., such as a Webcast). The user's request is received and processed by the redirection subsystem 610, which (at 715) directs the user to a particular POP site 620 from which the stream will be delivered (e.g., by returning a path directing the user's streaming application that POP site 620).

At 720, a load balancer module 625 residing at the POP site 620 selects a particular leaf splitter (e.g., splitter 633) from a group of leaf splitters 631-635 at the site. In one embodiment, the load balancer 625 is a layer 4 switch which continually monitors the load on each of the leaf splitters 631-635, and assigns the new user request to the least-loaded splitter. In this embodiment, the layer 4 switch may be identified by a virtual internet protocol ("VIP") address included in the path sent by the redirection subsystem 610.

sub A10

At 725, a root splitter failure is detected by the monitoring subsystem (e.g., via one or more of the failure detection techniques described above). As a result, the monitoring subsystem directs the backup root splitter 631 (e.g., via the backup agent 641) to reconfigure itself as the new primary root splitter 630. In addition, the monitoring subsystem and/or the backup agent 641 directs the load balancer 625 to remove the backup root splitter 631 from the group of leaf splitters 631-635 monitored by the load balancer module 625.

It should be noted that the new primary root splitter 631, load balancer 625, leaf splitters 632-625, and/or redirection subsystem 610 may be reconfigured differently following a root splitter 630 failure depending on the streaming formats supported by the system. When a user requests a file encoded in a RealPlayer® streaming format (e.g., through a RealPlayer application residing on the client computer) the redirection subsystem transmits a comprehensive path to the user, specifying the location of the streaming file and the servers through which the data stream will pass on its way to the user. For example, a path such as "rtsp:\\<VIP> \Split\<Root Server IP>\Split\ <Origin Server IP>\<Encoder IP> \live_stream.rm" may be passed to the client's streaming application, identifying the virtual IP address of the load balancer 625, the root splitter 630, the origin server 510, and the encoder 530 as well as the name of the actual streaming file ("live_stream.rm"). Accordingly, when the backup root server 631 is reconfigured as the primary root server 630 as described above, the path subsequently transmitted to users by the redirection subsystem 610 is updated to

reflect the new IP address of the root splitter in the <Root Server IP> field (i.e., the IP address of the former backup server 631).

By contrast, if the system is configured to support the Windows Media Technologies ("WMT") streaming format, each server within the streaming path may need to be updated following a root splitter 630 crash. More particularly, in this embodiment, WMT server "publishing points" are reconfigured beforehand on each server to point back to the upstream splitter 630, origin server 510, and/or encoder 530.

For example, following a user request for streaming content an exemplary URL returned by the Redirection Subsystem 610 may look something like: "mms://<VIP>/ <Broadcast Publishing Point(3)>." Each edge splitter behind the VIP address would then expose "Broadcast Publishing Point(3)" and would thereby be configured to reference, for example, "<Root Splitter IP>/<Broadcast Publishing Point(2)>." Continuing with this example, the WMT root splitter 630, in turn, would expose "Broadcast Publishing Point(2)" which would be pre-configured to reference "<Origin Server IP>/<Broadcast Publishing Point(1)>." Finally, "Broadcast Publishing Point(1)," exposed by the WMT origin server 510 in one embodiment, would be filled with the stream received from the streaming encoder 530.

Thus, in this embodiment, when the monitoring subsystem detects a failure in the primary WMT root splitter 630, it must not only reconfigure the load balancer 625 to remove the backup root splitter 631 from the leaf splitter group, it must also reconfigure all the "Broadcast Publishing Point(3)" on the remaining leaf splitters 632-635 so that they reference the new root splitter 631. In addition, the new root splitter 631 must be reconfigured to remove its "Broadcast Publishing Point(3)" and now expose "Broadcast Publishing Point(2)," which points back to the origin server's 510's publishing point.

Accordingly, in can be seen that one of the benefits of the foregoing configuration is that the Redirection Subsystem 610 can offer up the same URL to the user, even after the primary root splitter 630 fails. The user will then be provided the same stream using a newly-configured set of publishing points.

Regardless of which streaming format is used, new streams are provided to users through the new primary root splitter at 735. At 740, the operations staff at the data center is notified of the primary root server failure. The operations staff may then attempt to evaluate and solve the problem remotely before making a trip to the POP site 620.

Embodiments of the present invention include various steps, which have been described above. The steps may be embodied in machine-executable instructions. The instructions can be used to cause a general-purpose or special-

purpose processor to perform certain steps. Alternatively, these steps may be performed by specific hardware components that contain hardwired logic for performing the steps, or by any combination of programmed computer components and custom hardware components.

Elements of the invention may also be provided as a machine-readable medium for storing the machine-executable instructions. The machine-readable medium may include, but is not limited to, floppy diskettes, optical disks, CD-ROMs, and magneto-optical disks, ROMs, RAMs, EPROMs, EEPROMs, magnet or optical cards, propagation media or other type of media/machine-readable medium suitable for storing electronic instructions. For example, the present invention may be downloaded as a computer program which may be transferred from a remote computer (e.g., a server) to a requesting computer (e.g., a client) by way of data signals embodied in a carrier wave or other propagation medium via a communication link (e.g., a modem or network connection).

Throughout the foregoing description, for the purposes of explanation, numerous specific details were set forth in order to provide a thorough understanding of the invention. It will be apparent, however, to one skilled in the art that the invention may be practiced without some of these specific details. For example, although the foregoing embodiments were described in the context of specific streaming formats (e.g., Windows Media and RealMedia), various other streaming media formats may be implemented consistent with the

underlying principles of the invention. Accordingly, the scope and spirit of the invention should be judged in terms of the claims which follow.